Machine Learning (MC 321)
Lab Assignment 5

Question 1

Learn a Naïve-Bayes binary classifier on the dataset generated using the following data sampling:

1. Draw 50 random samples from $N(5, 20)$.
2. Draw 50 random samples from $N(11, 10)$.
3. Draw 50 random samples from $N(20, 8)$.

Consider the mean and variance of two classes as:

- Class 1: $\mu = 8, \sigma^2 = 20$
- Class 2: $\mu = 16, \sigma^2 = 25$

Classify using the Naïve-Bayes classifier:

- Assume *a priori* probabilities as $(0.5, 0.5)$, $(0.3, 0.7)$, and $(0.7, 0.3)$.
- Visualize data and class by plotting a histogram.
- Apply Laplace smoothing for computing the conditional probabilities:

$$P(X = x \mid \text{class} = c) = \frac{\text{number of examples with } X = x \text{ and class } = c + \alpha}{N + \alpha \cdot K}$$

  where N is the total number of examples from class c, K is the number of values taken by the feature X, and α is the smoothing parameter (typically $\alpha = 1$).

Question 2

Consider the Iris flower dataset that contains a set of 150 samples, which consists of 50 samples from each of three species of Iris: setosa (label 0), versicolor (label 1), and virginica (label 2). Each sample was measured in four features: sepal length, sepal width, petal length, and petal width.

Data Preparation and Visualization:

- Split the dataset into a balanced (with respect to the labels) training and test set, containing respectively 80% and 20% of the dataset.

Perceptron Learning: Consider the Iris flower dataset.

- Visualize the first two features of the training set, i.e., sepal length and sepal width, and their corresponding labels/classes.
- Now consider only the dataset containing two classes: setosa and versicolor.
- Classify the dataset into two classes with the Perceptron. Report the training and test errors. Comment.

Question 3

Perform binary classification using logistic regression on the data in the file *Social Network Ads.csv*, which is a categorical dataset to determine whether a user purchased a product or not by using three features.

1. Visualize the data by 3D plotting features using different colors for label 0 and 1.
2. Plot the training data, test data, and decision boundary learned by logistic regression in the same figure. (The boundary should be a straight line separating the region where $h_w(x) > 0.5$ from the region where $h_w(x) \leq 0.5$ $(h_w(x^{(i)}) = \sigma(w^T x^{(i)})))$ for the above three results.
3. Use 90% of the data points for training and the remaining 10% for testing the accuracy of classification.
4. Using the confusion matrix, find accuracy, precision, F1 score, and recall.